УДК 53.088, 519.23

МЕТОД СТАТИСТИЧЕСКОГО СРАВНЕНИЯ ДАННЫХ И ЕГО ПРИМЕНЕНИЕ ДЛЯ АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ ЯДЕРНО-ФИЗИЧЕСКИХ ДАННЫХ

С.И. Битюков, Н.В. Красников*, А.В. Максимушкина, А.Н. Никитенко***, В.В. Смирнова**

Институт физики высоких энергий, Протвино, Россия
**** Институт ядерных исследований РАН, Москва, Россия***
***** ИАТЭ НИЯУ МИФИ, Обнинск, Россия***
****** Империал колледж, Лондон, Великобритания***

Предлагается метод статистического сравнения данных для использования в задачах анализа экспериментальных и моделированных ядерно-физических данных, являющийся развитием метода статистического сравнения гистограмм. Подробно описан алгоритм процедуры сравнения данных. В качестве меры различия данных используется двумерная тест-статистика, определяемая через статистические моменты распределения, полученного при вычислении значимостей различия измеренных и (или) смоделированных значений определяемой величины в соответствующих точках измерения. Значимости различия реализованных значений измеряемой величины в каждой точке являются реализацией случайной величины, распределение вероятностей которой близко к стандартному нормальному распределению, если измеряется та же самая величина и в методах измерения отсутствуют систематические различия. Это позволяет избежать зависимости результатов от формы распределения. Точность оценки меры различия определяется с помощью эксперимента Монте-Карло. Показана возможность применения предлагаемого метода в задачах сравнения данных разных экспериментов или экспериментальных и смоделированных данных.

Ключевые слова: теория распределений и методы Монте-Карло, измерения и теория ошибок, анализ данных (алгоритмы и применение).

ВВЕДЕНИЕ

Развитие ядерной энергетики является неотъемлемой частью развития современной цивилизации. Устойчивый интерес к ядерным реакциям привел к созданию как специализированных баз данных, например, EXFOR [1], так и специализированных программных продуктов, например, CASCADE/INPE [2] или CASCADeX [3], позволяющих моделировать ядерные реакции в соответствии с заданными модельными предположениями. Сравнение существующих экспериментальных данных и данных, полученных моделированием ядерных реакций, является одной из основных задач при выборе теоретических построений, адекватно описывающих экспериментальные данные. Часто при сравнении данных используются факторы согласия [4] или их

взвешенная комбинация [3] с не совсем очевидной интерпретацией полученных результатов сравнения данных.

В работах [5, 6]рассматривается метод статистического сравнения гистограмм как возможное решение задачи сравнения данных унифицированным способом. Преимущество данного подхода в сравнении со стандартным методом, использующим хи-квадрат-распределение, показано в работе [7]. Работы [5 – 7] ориентированы на сравнение потоков выборочных данных, однако основные идеи этих работ можно использовать и при сравнении наборов данных произвольной природы.

Итак, предположим, что нужно сравнить два набора данных $h1$и $h2$, состоящих из M измеренных в точках $x_1, x_2, \ldots, x_M$ значений

$$h1: \quad \hat{n}_{11} \pm \hat{\sigma}_{11}, \hat{n}_{21} \pm \hat{\sigma}_{21}, \ldots \hat{n}_{M1} \pm \hat{\sigma}_{M1}$$

и

$$h2: \quad \hat{n}_{12} \pm \hat{\sigma}_{12}, \hat{n}_{22} \pm \hat{\sigma}_{22}, \ldots \hat{n}_{M2} \pm \hat{\sigma}_{M1} \ .$$

Сравнив эти два набора данных, нужно принять решение о том, описывают ли эти данные одно и то же явление (т.е. результаты получены из одной и той же генеральной совокупности) или результаты соответствуют разным явлениям (например, измеряются разные процессы или не учтены какие-нибудь особенности собственно измерений, т.е. результаты взяты из разных генеральных совокупностей), а также оценить вероятность того, что решение об их различии правильное.

РАССТОЯНИЕ МЕЖДУ НАБОРАМИ ДАННЫХ

Большинство методов сравнения данных используют в качестве меры различимости данных некоторое «расстояние между данными». Так, например, в методе χ^2 расстояние между двумя наборами данных, представленными в виде гистограмм [8], это

$$\chi^2 = \sum_{i=1}^{M} [(\hat{n}_{i1} / N_1 - \hat{n}_{i2} / N_2)^2 / (\hat{n}_{i1} / N_1^{\ 2} + \hat{n}_{i2} / N_2^{\ 2})] = \sum_{i=1}^{M} \hat{S}_i^2 \ ,$$

где $\hat{S}_i$ в случае пуассоновских потоков событий можно назвать «нормализованной значимостью различия» значения в i-ом канале в первой гистограмме и i-ом канале во второй гистограмме.

Отметим, что в предлагаемом методе также используется $\hat{S}_i$, но несколько иначе, чем в методе хи-квадрат. Существуют и другие «расстояния» (см. [8]).

РАСПРЕДЕЛЕНИЕ ТЕСТОВЫХ СТАТИСТИК

Предлагается использовать статистические моменты распределения $\hat{S}_i$, где i=1, …, M. Это распределение, состоящее из M значений, в случае, если оба набора данных получены из одной и той же генеральной совокупности, близко к стандартному нормальному распределению, поскольку каждая реализация случайной величины «значимость различия» значений для каждой точки измерения i является реализацией стандартной нормальной величины.

Таким образом, в качестве расстояния между наборами данных предлагается не одномерная величина, как в других методах, а многомерная. В рассмотренных далее примерах двумерная величина

$$SRMS = (\overline{S}, \ RMS), \quad \text{где} \quad \overline{S} = \sum_{i=1}^{M} \hat{S}_i / M \ ,$$

есть среднее значение распределения «нормализованных значимостей различия», а величина

$$RMS = \sqrt{\sum_{i=1}^{M} (\hat{S}_i - \overline{S})^2 / M}$$

является среднеквадратическим отклонением этого распределения.

SRMS имеет ясную интерпретацию:

– при *SRMS* = (0,0) оба набора данных идентичны;

– если *SRMS* ≈ (0,1), то оба набора данных получены из одной и той же генеральной совокупности;

– если вышеупомянутые условия не выполняются, то наборы данных получены из разных генеральных совокупностей.

Отметим, что существует взаимосвязь между средним, среднеквадратическим и значением хи-квадрат:

$$RMS^2 = \chi^2 / M - \overline{S}^2, \quad \chi^2 = \sum_{i=1}^{M} \hat{S}_i \ .$$

Данная взаимосвязь указывает на то, что тест-статистика χ^2 является комбинацией двух тест-статистик – *RMS* и $\overline{S}$.

ЗНАЧИМОСТЬ РАЗЛИЧИЯ

Рассмотрим модель, в которой наборы данных *h*1 и *h*2 определены следующим образом: случайная переменная *n* (измеряемая величина в точке измерения *i*) подчиняется нормальному распределению

$$\varphi(n|n_{ik}) = \exp[-(n - n_{ik})^2 / (2\sigma_{ik}{}^2)] / (\sqrt{2\pi}\sigma_{ik}).$$

Здесь ожидаемое значение в *i*-ой точке измерения *k*-го набора данных есть n_{ik}, а ожидаемая величина дисперсии – $\sigma_{ik}{}^2$.

Для случая сравнения двух наборов данных введем значимость различия в соответствующих точках измерения

$$\hat{S}_i = (\hat{n}_{i1} - \hat{n}_{i2}) / \sqrt{\hat{\sigma}_{i1}{}^2 + \hat{\sigma}_{i2}{}^2} \ .$$

В данном случае $\hat{n}_{ik}$ – это наблюденное значение в точке измерения *i* набора данных *k*, – соответствующее стандартное отклонение. Если предположить, что во втором наборе данных нет статистических ошибок, то значимость различия будет равна

$$\hat{S}_i = (\hat{n}_{i1} - \hat{n}_{i2}) / \hat{\sigma}_{i1} \ .$$

ГЕНЕРАЦИЯ ПОВТОРНОГО НАБОРА ДАННЫХ

Следующий шаг является важным в данном методе сравнения наборов данных. По аналогии с генерацией повторной выборки в методе «бутстреп» [9] он может быть назван генерацией повторного набора данных. Для каждого из сравниваемых наборов данных создается определенное количество подобных наборов данных (клонов) в соответствии с рассматриваемой моделью, а именно, значение в каждой точке измерения клонируемых наборов данных разыгрывалось в соответствии с законом $\mathbb{N}(\hat{n}_{ik}, \hat{\sigma}_{ik})$. Это позволяет создать две имитационные модели генеральных совокупностей наборов данных для сравниваемых наборов данных. В рассмотренных ниже примерах было смоделировано 4999 клонов для каждого из наборов данных и затем проведено 5000 сравнений пар полученных наборов данных. В ходе каждого

сравнения строилось распределение значимостей различия в соответствующих точках измерения и определялось среднее и среднеквадратическое полученного распределения. Полученные величины используются для проверки гипотезы о принадлежности наборов данных одной или разным генеральным совокупностям, определения ошибок первого (α) и второго (β) рода и оценки вероятности правильности решения о том, что наборы данных принадлежат разным генеральным совокупностям.

РАЗЛИЧИМОСТЬ НАБОРОВ ДАННЫХ

Обычно для различимости гипотез при сравнении наборов данных задают уровень значимости критерия, т.е. вероятность совершить ошибку первого рода α, и вычисляют мощность критерия $1-\beta$, где β – вероятность ошибки второго рода. Различимость наборов данных можно также оценить с помощью некоторой функции ошибок первого (α) и второго (β) рода, которая фактически является вероятностью правильного заключения о принадлежности наборов данных разным генеральным совокупностям. Если эта величина равна единице, то наборы данных 100%-но различимы. Если же эта величина равна нулю, то наборы данных неразличимы, и можно утверждать, что они взяты из одной генеральной совокупности.

Если критическая область (величина, линия, поверхность и т.п.) выбрана корректно, т.е. выполнено условие $\alpha + \beta \leq 1$, то вероятность правильного решения о различимости наборов данных определяется как $1 - (\alpha+\beta)/(2-(\alpha+\beta))$ [10].

ПРИМЕРЫ

Получение экспериментальных ядерно-физических данных является сложной задачей, поэтому для расчетов широко используются данные (например, данные по сечениям реакций), полученные с помощью специализированных пакетов программ, в основе которых лежат физические модели, например, [11]. Таким образом, важной задачей является определение, насколько корректно расчетные данные описывают эксперимент.

Использование метода позволяет не только определить принадлежность экспериментальных и расчетных данных одной генеральной совокупности (когда модель хорошо описывает эксперимент), но и найти доверительные интервалы параметров расчетной модели, в пределах которых модель хорошо согласуется с экспериментом. Также можно определить, являются ли различия в экспериментальных данных, полученных на разных установках или различными группами экспериментаторов, настолько серьезными, что эти данные не могут быть объединены для совместного рассмотрения.

Рассмотрим несколько наборов данных и проведем их анализ с помощью разработанного метода.

Данные по сечениям реакции ^{52}Cr (p,x)^{7}Be. Экспериментальные данные брались из EXFOR [1], а расчетные были получены с помощью новой версии программы CASCADE [3]. Исходные данные представлены в табл. 1 и на рис. 1.

В результате обработки этих данных были получены двумерные распределения величин *S* и *RMS* для соответствующего распределения значимостей различия, показанные на рис. 2, где критическая линия для проверки гипотезы о разделимости наборов данных отделяет полупроцентный уровень значимости (ошибка первого рода $\alpha = 0.005$) от нижнего распределения (калибровочного) и позволяет вычислить мощность критерия.

В рассмотренном примере $\alpha = 0.005$ и $\beta = 0.0202$. Решение о том, что данные взяты из разных генеральных совокупностей, верно с вероятностью 0.99746.

Таблица 1

Исходные данные

E_p, МэВ	Расчетные сечения, мб	Экспериментальные сечения, мб	Ошибка эксперимента, мб
50	0.0622	0.056	0.01
100	0.183	0.297	0.028
150	0.342	0.348	0.032
250	0.6835	0.676	0.069
400	1.292	1.26	0.11
600	2.962	2.39	0.2
800	3.824	3.28	0.33
1200	5.525	5.38	0.46
1600	6.84	6.04	0.54
2600	8.358	6.58	0.61

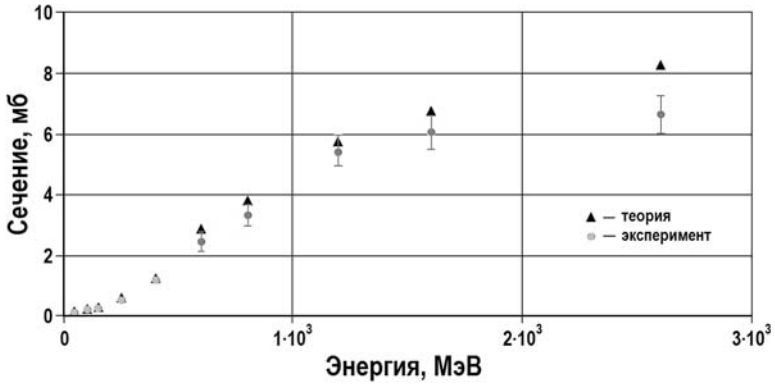

Рис. 1. Теоретические и экспериментальные данные реакции ^{52}Cr (p,x)^{7}Be

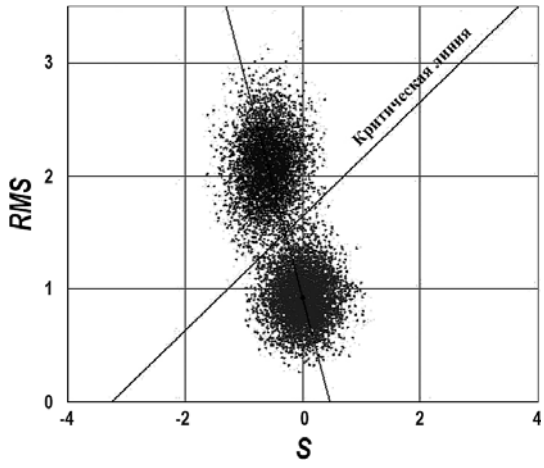

Рис. 2. Оценка различимости наборов данных

Сечения для реакции $^{56}Fe(p,x)^{24}Na$. Экспериментальные сечения взяты из EXFOR[1], а два набора расчетных сечений получены с помощью программы CASCADE [3]. Один набор расчетных сечений был получен с использование модели образо-

вания кластеров с $A > 10$, а второй – без учета их образования (табл. 2).

Таблица 2

Сечения для реакции $^{56}Fe(p,x)^{24}Na$

E_p, МэВ	Экспериментальные сечения, мб	Ошибка эксперимента, мб	Расчетные сечения по первой модели, мб	Расчетные сечения по второй модели, мб
200	1.29E-02	3.00E-04	3.6120E-02	3.6123E-03
300	4.84E-02	2.00E-04	1.4450E-02	6.0209E-03
585	2.66E-01	8.50E-02	7.5850E-01	1.5223E-01

Результаты использования метода показаны на рис. 3.

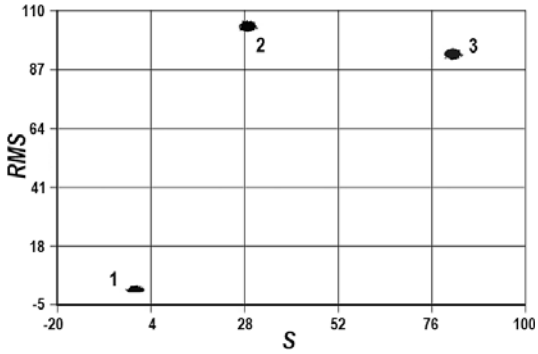

Рис. 3. Пятно 1 (калибровочное) соответствует экспериментальным данным; пятно 2 – результат сравнения данных для модели с учетом образования кластеров и экспериментальных данных; пятно 3 – результат сравнения модели без учета образования кластеров с экспериментальными данными

В этом примере расчетные данные для двух моделей не принадлежат одной генеральной совокупности с экспериментальными данными (т.е. не могут быть использованы при описании экспериментальных сечений). Такое расхождение можно объяснить тем, что точность моделирования эксперимента не позволяет описать точные измерения либо приводимые ошибки измерений не соответствуют реальной точности полученных данных.

Данные двух экспериментальных групп – сечения реакции $^{27}Al(p,x)^{7}Be$. Сечения получены двумя группами экспериментаторов (группы Михеля и Титаренко) [1].

Таблица 3

Сечения для реакци$^{27}Al(p,x)^{7}Be$

E_p, МэВ	Экспериментальные сечения Михеля, барн	Ошибка экспериментальных сечений Михеля, барн	Экспериментальные сечения Титаренко, барн	Ошибка экспериментальных сечений Титаренко, барн
40,8	0,000187	2,63E-05	0,00019	4,00E-05
44,6	0,000318	5,58E-05	0,00032	7,00E-05
66	0,000635	3,41E-05	0,00065	6,00E-05
67	0,000649	3,36E-05	0,00076	0,00021
600	0,00502	0,000355	0,00401	0,00022
799	0,00659	0,000474	0,00654	0,00033
800	0,00644	0,000457	0,0064	0,0004
1200	0,00848	0,00063	0,0083	0,0006
1600	0,00865	0,00063	0,00865	0,0004
2600	0,0095	0,00072	0,0092	0,0007

Результаты анализа с использованием метода статистического сравнения данных показаны на рис. 4.

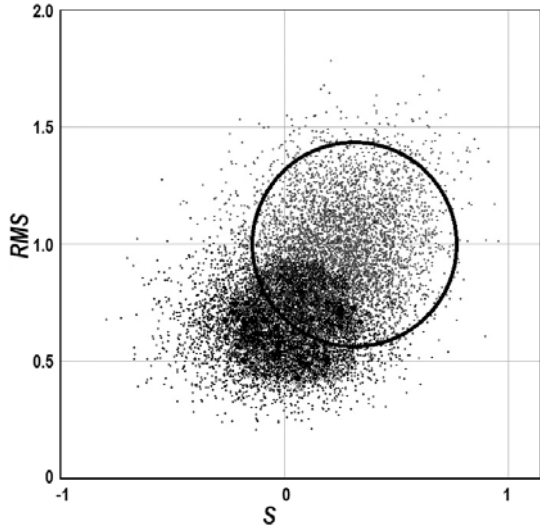

Рис. 4. Черное пятно (внизу) – калибровочное распределение по данным Михеля, сравниваемым со своими клонами через генерацию повторных наборов данных; темно-серое пятно (границы выделены эллипсом, частично перекрывается с калибровочным) – результат сравнения данных Титаренко и Михеля

На рисунке наблюдается существенное перекрытие распределений. Таким образом, объединение экспериментальных данных этих групп требует дополнительных исследований.

ЗАКЛЮЧЕНИЕ

Предложенный подход с применением двумерной тест-статистики позволяет значительно усилить мощность критерия при проверке гипотез по сравнению с методами, использующими одномерные тест-статистики.

Метод может быть использован для сравнения данных, полученных в различных экспериментах, а также экспериментальных и моделированных ядерно-физических данных.

Работа выполнена при частичной поддержке РФФИ (грант 13-02-00363).

Литература

1. EXFOR Library. Available: http://www-nds.iaea.org/exfor/exfor.htm

2. *Барашенков В.С., Конобеев А.Ю., Коровин Ю.А., Соснин В.Н.*//Атомная энергия. – 1999. - №87. - С. 283.

3. *Андрианов А.А., Конобеев А.Ю., Коровин Ю.А., Купцов И.С., Станковский А.Ю.*//Известия вузов. Ядерная энергетика. – 2011. - №2.– С. 5.

4. *Коровин Ю.А., Максимушкина А.В.*//Ядерная физика и инжиниринг. – 2014. - №5. – С. 237.

5. *Bityukov S.I., Krasnikov N.V., Nikitenko A.N., Smirnova V.V.* A method for statistical comparison of histograms arXiv:1302.2651 - 2013.

6. *Bityukov S.I., Krasnikov N.V., Nikitenko A.N., Smirnova V.V.*//ВестникРУДН.Серия: математика, информатика,физика- 2014. - №2 – С. 324.

7. *Bityukov S., Krasnikov N., Nikitenko A., Smirnova V.*//Eur.Phys.J.Plus- 2013.- №128:143.

8. *PorterF.* Testing consistency of two histograms arXiv:0804.0380— 2008.

9. *Efron B.* Bootstrap methods: another look at the jackknife/ / Annals of Statistics- 1979 – 7.P. 1.

10. *Bityukov S.I., Krasnikov N.V.*, Distinguishability of Hypotheses// Nucl.Inst.&Meth. – 2004-A534. P.152.

11. *Коровин Ю.А., Максимушкина А.В.//*Известия вузов. Ядерная энергетика. – 2014. - №2. – С. 51.

Поступила в редакцию 09.12.2013 г.

Авторы

Битюков Сергей Иванович, ведущий научный сотрудник, доктор физ.-мат. наук
E-mail: Serguei.Bitioukov@cern.ch

Красников Николай Валерьевич, заведующий отделом теоретической физики
доктор физ.-мат. наук
E-mail: Nikolai.Krasnikov@cern.ch

Максимушкина Анастасия Владимировна, ассистент
E-mail: a.v.saenko@mail.ru

Никитенко Александр Николаевич, научный сотрудник, кандидат физ.-мат. наук
E-mail: Alexsandre.Nikitenko@cern.ch

Смирнова Вера Васильевна, старший научный сотрудник, кандидат физ.-мат. наук
E-mail: Vera.Smirnova@ihep.ru

UDC 53.088, 519.23

A METHOD FOR STATISTICAL COMPARISON OF DATA SETS AND ITS USES IN ANALYSIS OF NUCLEAR PHYSICS DATA

Bityukov S.I., Krasnikov N.V.*, Maksimushkina A.V.**, Nikitenko A.N.***, Smirnova V.V.

Institute for High Energy Physics, Protvino, Russia

* INR RAS, Moscow, Russia

** Obninsk Institute for Nuclear Power Engineering, National Nuclear Research University «MEPhI». 1 Studgorodok, Obninsk, Kaluga reg., 249040 Russia

*** Imperial College, London, UK

ABSTRACT

We propose a method for statistical comparison of two data sets. The method is based on the method of statistical comparison of histograms. Usually a one-dimensional test statistic is used as a measure of distinction of data sets. This test statistic depends on the shape of distributions in data sets. Using the two-dimensional test statistics which is determined via the statistical moments of distribution produced by the calculation of "the significance of deviations" for the corresponding points with observed values is proposed in the paper as a distinction measure between data sets. The significance of deviation in the corresponding points can be considered as a realization of the random variable which is close to a standard normal random variable if we observe the same random value in both data sets. It helps to avoid the dependence of the result on the shape of distributions. The accuracy of the estimator for the measure of distinction is determined by the Monte-Carlo experiment which, by analogy with the construction of repeated samples (resampling) in the bootstrap method, it is possible to call construction of repeated data set (redatasetting). As an estimator of quality of the decision made, it is proposed to use the value which it is possible to call the probability that the decision "data sets are various" is correct.

Key words: distribution theory and Monte-Carlo studies, measurement and error theory, data analysis (algorithms and implementation).

REFERENCES

1. EXFOR Library. Available at: http://www-nds.iaea.org/exfor/exfor.htm

2. Barashenkov V.S., Konobeev A.Yu., Korovin Yu.A., Sosnin V.N. *Atomnaya Energiya*. 1999,

no. 87, p. 283.

3. Anrianov A.A., Konobeev A.Yu., Korovin Yu.A., Kuptsov I.S., Stankovskij A.Yu. *Izvestiya vuzov. Yadernaya energetika*. 2011, no. 2, p. 5.

4. Korovin Yu.A., Maksimushkina A.V. *Yadernaya Fizika i Ingeniring*. 2014, no. 5, p. 237.

5. Bityukov S.I., Krasnikov N.V., Nikitenko A.N., Smirnova V.V. *A method for statistical comparison of histograms*. arXiv:1302.2651 - 2013.

6. Bityukov S.I., Krasnikov N.V., Nikitenko A.N., Smirnova V.V. *Vestnik RUDN. Seriya: matematika, informatika, fisika*. 2014, no.2, p. 324.

7. Bityukov S., Krasnikov N., Nikitenko A., Smirnova V. *Eur. Phys. J. Plus*. 2013, no. 128:143.

8. Porter F. *Testing consistency of two histograms*. arXiv:0804.0380 - 2008.

9. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*. 1979, no. 7, p. 1.

10. Bityukov S.I., Krasnikov N.V. Distinguishability of Hypotheses. *Nucl. Inst. &Meth*. 2004-A534, p. 152.

11. Korovin Yu.A., Maksimushkina A.V. *Izvestiya vuzov. Yadernaya energetika*. 2014, no. 2, p. 51.

Autors

Bityukov Sergey Ivanovich, Leading Scientist, Dr. Sci. (Phys.-Math.)
E-mail: Serguei.Bitioukov@cern.ch

Krasnikov Nikolai Valer'evich, Head of Theoretical Physics Department, Professor, Dr. Sci. (Phys.-Math.)
E-mail: Nikolai.Krasnikov@cern.ch

Maksimushkina Anastasiya Vladimirovna, Assistant
E-mail: a.v.saenko@mail.ru

Nikitenko Alexandr Nikolaevich, Research Associate, Cand. Sci. (Phys.-Math.)
E-mail: Alexsandre.Nikitenko@cern.ch

Smirnova Vera Vasil'evna, Senior Research Scientist, Cand. Sci. (Phys.-Math.)
E-mail: Vera.Smirnova@ihep.ru